Available online at www.sciencedirect.com
ScienceDirect
Genomics 90 (2007) 493–501
GENOMICS
www.elsevier.com/locate/ygeno

Identification, evolution, and association study of a novel promoter and first exon of the human *NOD2* (*CARD15*) gene☆

Kathy King [a], Richard Bagnall [a], Sheila A. Fisher [a], Faisal Sheikh [a], Andrew Cuthbert [a], Sipin Tan [a], Nicholas I. Mundy [b], Philip Rosenstiel [c], Stefan Schreiber [c], Christopher G. Mathew [a], Roland G. Roberts [a,*]

[a] *Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London SE1 9RT, UK*
[b] *Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK*
[c] *Institute for Clinical Molecular Biology, Christian-Albrechts-University Kiel, 24105 Kiel, Germany*

Received 15 June 2007; accepted 19 July 2007
Available online 24 August 2007

Abstract

Mutations in the *NOD2* (*CARD15*) gene predispose to Crohn's disease (CD), a human chronic inflammatory bowel disorder, and can cause Blau syndrome. During an investigation of an apparent correlation between a frameshifting mutation in the canonical first exon of *NOD2* of marmoset and tamarin species and their susceptibility to chronic colitis, we found that, contrary to previous reports, the basal levels of *NOD2* transcripts in tissues relevant to CD arise from a distinct novel promoter and first exon. The canonical first exon, by contrast, seems to be of negligible transcriptional importance under physiological conditions, and its reading frame has been disrupted twice during primate evolution. Thus the main NOD2/CARD15 protein isoform produced in humans and other primates is 27 amino acids shorter than previously reported, starting at a conserved methionine in exon 2. We show that there is no significant association between variants in the novel *NOD2* promoter region and CD.

Keywords: Crohn's disease; NOD2; CARD15; Blau syndrome; Promoter regions; Colitis; Callitrichinae; Marmoset

Crohn's disease (CD) is a form of inflammatory bowel disease characterized by chronic relapsing ulceration and inflammation of the gastrointestinal tract [1]. Genetic studies identified the *NOD2* gene (also called *CARD15*) as a susceptibility gene for CD [2–4]. Three commonly observed variants in *NOD2* (R702W, G908R, and L1007fs) are associated with CD, although multiple rarer variants have been identified, some of which are also likely to predispose to CD [5–7]. Heterozygotes for these mutations have a 2- to 3-fold increase in risk of CD, with a 17-fold increase in homozygotes or compound heterozygotes [8]. *NOD2* activates NF-κB in response to the peptidoglycan component of bacterial cell walls via its ligand muramyl dipeptide, and *NOD2* mutations result in a reduction of this response [9,10]. Thus a constitutional deficiency in the innate immune response appears to contribute to the pathogenesis of CD. *NOD2* is also mutated in the rare autosomal dominant disorder Blau syndrome, but the mutational spectrum is different from that of CD and is associated with a gain rather than a loss of function [11]. The *NOD2* gene, which is located on chromosome 16q21, has 12 exons and encodes a protein of 1040 amino acids [12]. The protein contains two N-terminal caspase recruitment domains (CARDs), a central nucleotide-binding oligomerization domain (NOD), and multiple C-terminal leucine-rich repeats. An isoform of NOD2 (NOD2-S) that is generated by skipping of exon 3 and is truncated in the second CARD can down-regulate the activation of NF-κB by NOD2 [13]. Multiple splice variants of *NOD2* have been reported, the functional significance of which is unknown [14].

In our previous study of the evolution of *NOD2* we found that the open reading frame (ORF) of humans and apes appears to start in the first exon [6], contributing an additional N-

☆ Sequence data from this article have been deposited the EMBL/GenBank Data Libraries under Accession Nos. DQ868963–DQ868975.

* Corresponding author. Fax: +44 20 7188 2585.
E-mail address: roli.roberts@genetics.kcl.ac.uk (R.G. Roberts).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.07.009

terminal 27 amino acids, which are not present in other vertebrates. We also showed that the New World monkey *Saguinus oedipus,* which is known to be highly susceptible to chronic colitis [15], has a frameshifting mutation in the canonical exon 1 of *NOD2.* This was of considerable interest, since it raised the possibility that a lack of expression of *NOD2* might contribute to susceptibility to colitis in this species. The initial characterization of *NOD2/CARD15* [12] revealed two potential in-frame translation initiation sites at Met1 and Met28, both of which were used, although Met1 appeared to be more efficient in an in vitro translation assay [12]. The presence of these alternative initiation sites, combined with the nonorthology of the first exon of *NOD2* in the three independently deposited *NOD2* cDNA sequences (human, mouse, and cow), has led us to explore the biological significance of the proposed first exon of the human gene. We show here that a novel alternative promoter and novel first exon of *NOD2* are responsible for the abundant constitutive transcript isoforms in human tissues and that primate NOD2, like other vertebrate NOD2's, is likely to be translated from the first methionine codon in exon 2 (known as Met28) to produce a protein of 1023 rather than 1040 amino acids. We have also evaluated the novel promoter and first exon for sequence variants that might predispose to Crohn's disease.

Results

Canonical exon 1 ORF has been disrupted twice during primate evolution

While acquiring sequences of *NOD2* exons from various primate species, we noted that the New World monkey *S. oedipus* (cotton-top tamarin; Accession No. AY594138), but not human, chimpanzee, or gibbon (Accession Nos. AF178930, AY594162, and AY594150), had a frameshifting insertion of a single G nucleotide in canonical exon 1 [6] (hereafter referred to as CanE1), meaning that, as in nonprimates such as mice and cattle, translation would have to commence in exon 2. Recalling that *S. oedipus* (in captivity, but not in the wild) is known to have a species-wide predisposition to a chronic colitis that has been likened to Crohn's disease [15–17], it occurred to us that this insertion might be a candidate predisposing mutation. We therefore used genomic PCR and direct sequencing to acquire CanE1 sequences from a range of other callitrichid monkeys (the emperor tamarin *S. imperator,* the red-bellied tamarin *S. labiatus,* the saddle-back tamarin *S. fuscicollis,* Goeldi's monkey *Callimico goeldii,* the common marmoset *Callithrix jacchus,* and the golden-headed lion tamarin *Leontopithecus chrysomelas;* Accession Nos. DQ868965–DQ868971), as well as a cebid New World monkey (owl monkey *Aotus trivirgatus;* Accession No. DQ868972). These sequences showed that all callitrichids had the same insertional mutation (Fig. 1) but that the non-callitrichid *A. trivirgatus* had an intact open reading frame, suggesting that this apparent frameshift was a callitrichid-specific variant [18].

We studied the literature for evidence of bowel disorders in callitrichids other than *S. oedipus* and duly found numerous peer-reviewed and anecdotal reports of high incidence of chronic colitis in *S. mystax, S. labiatus, S. fuscicollis, S. oedipus, Ca. jacchus* (often labeled as "wasting marmoset syndrome"), but not in non-callitrichid New World monkeys [19–25] (unlike *S. oedipus,* colitis in other callitrichids is not associated with a high incidence of colon cancer). This apparently perfect correlation between a disrupted *NOD2* open reading frame and a high prevalence of chronic colitis in captivity was clearly tantalizing, and led us to test the effects of such a mutation on the translation of the *NOD2* transcript; Ogura et al. have previously shown that methionines in both CanE1 (Met1) and exon 2 (Met28) may be used for translation initiation (their Fig. 3D) [12], so the likely translational consequences were unclear.

We therefore generated a series of constructs in the mammalian expression vector pGW1, fusing various versions of *NOD2* CanE1 with *NOD2* exon 2 and sequence encoding a readily detectable C-terminal hemagglutinin (HA) tag (see Supplementary Fig. 1A). These constructs included normal sequences from *Homo sapiens*, *S. oedipus,* and *A. trivirgatus,* together with versions of the human sequence with either Met1 or Met28 mutated to a valine ("M1V" and "M28V," respectively). These constructs were transfected transiently into COS-7 cells and the protein products were separated by high-resolution SDS–PAGE and detected by Western blotting with anti-HA antibody. In contrast to Ogura et al. [12], only products initiating from Met1 were seen, and these were generated only from those constructs bearing an intact upstream open reading frame (normal and M28V *H. sapiens* and normal *A. trivirgatus* sequences) and not from M1V *H. sapiens* or normal *S. oedipus* sequence (see Supplementary Fig. 1B). This suggests that Met28 cannot function as an efficient initiation codon in this context and that the callitrichid-specific insertion appears to be a translationally null mutation.

At this point we chanced to sequence *NOD2* CanE1 from the COS-7 cells that we had been using as a transfection host (Accession No. DQ868964). COS-7 cells are derived from *Cercopithecus aethiops* (African green monkey), which, like the great apes, is from the catarrhine branch of the primates. Surprisingly, this animal, which does not to our knowledge have a high incidence of colitis, was also found to have a frameshifting insertional mutation in CanE1, this time of two G nucleotides, at the same position as that found in the callitrichids (Fig. 1). The same dinucleotide insertion was found in the public sequence for *Macaca mulatta* (rhesus macaque), another Old World monkey. This result suggested that perhaps the primate-specific upstream open reading frame was not subject to purifying selection and was essentially free to vary. Indeed, the pairwise K_a/K_s values between CanE1 sequences of those primates that have an intact ORF are greatly in excess of unity (data not shown), suggesting absence of purifying selection, though the small size of exon 1 makes this result weak. Furthermore, the recently released genome sequence of the gray mouse lemur *Microcebus murinus* (a strepsirrhine) shows that the CanE1 sequence of this animal does not have an intact donor splice site (see Fig. 1A). In combination with our finding that Met28 could not drive efficient translation, at least

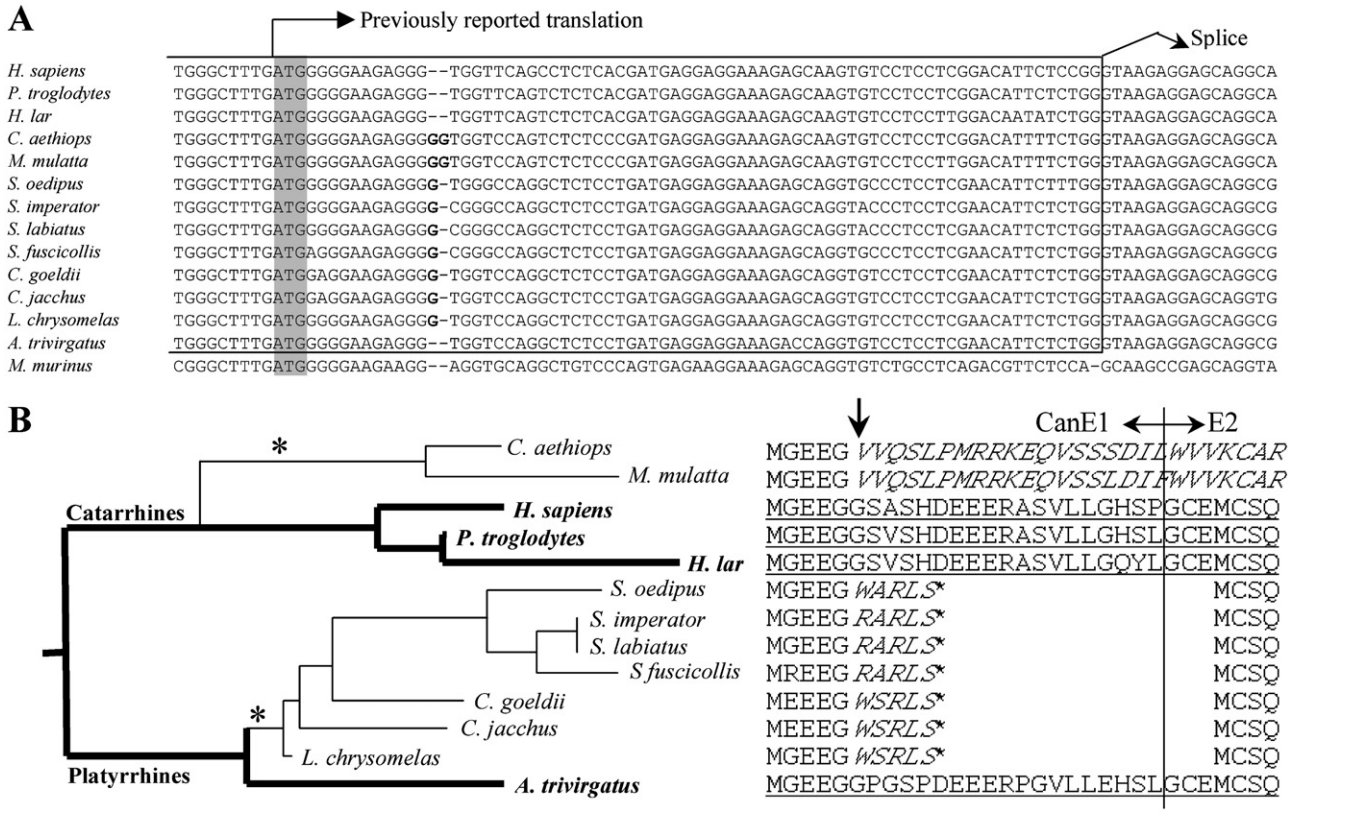

Fig. 1. *NOD2* canonical exon 1 (CanE1) ORF is not maintained in primates. (A) Alignment of the 3′ half of CanE1 from a range of primates. Box indicates alleged exon. Arrow indicates reported [12] translation initiation codon (shaded). Bold letters indicate insertion mutations in callitrichids and Old World monkeys. Accession Nos. DQ868963–DQ868972. (B) Maximum parsimony (DNAPars) phylogenetic tree (left-hand side) generated from full sequence of *NOD2* CanE1, rooted using *Mi. murinus* CanE1 as an outgroup and presented adjacent to the translations of the corresponding transcripts (right-hand side). The tree confirms the monophyly of the callitrichid sequences and reveals the most parsimonious points at which insertion mutations may have occurred (asterisks). Bold lines, bold species names, and underlined peptide sequence—putative intact ORF. Italics—shifted open reading frames. Vertical arrow—position of insertion mutations.

when positioned 3′ of CanE1, these data caused us to reassess the evidence that CanE1 actually *is* the first exon of primate *NOD2* genes.

Identification of a candidate alternative human NOD2 exon 1

The identification of the human *NOD2* CanE1 (as presented in GenBank cDNA entries NM_022162, AY187243, AY187246, and AF178930) originates in a paper by Ogura et al., in which 5′RACE was performed on RNA from a tissue pathogenetically less relevant to CD (mammary gland) [12]. Although subsequent workers have successfully performed RT-PCR between CanE1 and other exons [13,14,26], to our knowledge the use of this CanE1 has not been independently assessed. We decided to use bioinformatic methods to address its relationship to genomic and expressed sequences in other species.

The mouse *Nod2* exon 1 appears to be controversial—one cDNA sequence (AF520774) has a first exon that consists almost entirely of a MaLR LTR repeat, while the other (AY160221) has a unique GC-rich first exon. The former is not represented at all in the EST database, while the latter is present in 35 murine ESTs. We therefore assume that the sequence represented by AY160221 contains the major mouse *Nod2* exon 1. The only other independently acquired mammalian *Nod2* cDNA sequence (i.e., not merely predicted from the human) was bovine *NOD2,* AY518737.

Of particular concern was the fact that the alleged first exons of mouse, human, and bovine *NOD2* do not resemble each other. Their nonorthology is proven by the presence upstream of the human *NOD2* exon 2 of separate orthologues of the murine and bovine first exons (such that the alleged first exons are in the order murine–human–bovine; the orthologue of the canonical human first exon is also apparent in the mouse *Nod2* gene). As the interspecific similarity of these sequences is very low (small islands of similarity, each of marginal statistical significance, but colinearly arranged), the sequence comparison was aided by dot-plotting, which essentially adduces syntenic information to make orthology more apparent. While the sequence of the first exons of genes can often vary significantly between mammalian species, it is rare for them to be nonorthologous.

Naturally it is formally possible that the mammalian *NOD2* gene has multiple promoters, with different promoters being identified by chance in each of the three species. However, the absence of an orthologue of the bovine exon 1 in mouse, and its lack of a donor splice site in human, led us to assess the candidacy of the mouse exon 1 as a mammal-wide *NOD2* exon

1; moreover, it is the only *NOD2* first exon represented in the EST database. Strikingly, we found orthologues of mouse exon 1 in all mammalian genome sequences examined, and although there was substantial sequence variation within the exon, in each case the donor splice site and the surrounding 20–30 bp were highly conserved (see Fig. 2). To our mind this made this sequence, which we term "alternative exon 1" (hereafter, AltE1), a likely first exon for mammalian *NOD2* genes.

In addition to AltE1 sequences gleaned from the public databases, we used genomic PCR to amplify orthologous sequences from the Old World monkey *Ce. aethiops* and the New World monkey *S. oedipus* (Accession Nos. DQ868974 and DQ868975); these are also presented in Fig. 2B.

AltE1 is the principal NOD2 exon 1 in primates

We set about assessing the use of AltE1 in two ways. The first was to use 5′cRACE to survey sequences openendedly upstream of *NOD2* exon 2 in endogenous human transcripts. The second was to use semiquantitative RT-PCR to compare its incorporation into *NOD2* transcripts with that of CanE1.

5′cRACE is a means of amplifying unknown sequences 5′ of known sequences in RNA molecules [27]. It is achieved via a circular ssDNA intermediate, which allows true nesting of the subsequent PCR with a concomitantly high yield and specificity, even on low-abundance transcripts. We applied this to human total white blood cell *NOD2* transcripts, using primers positioned in exon 2. A broad band of 5′cRACE product was obtained; digestion with the rare-cutting restriction enzyme *Bss*HII, which cuts uniquely in AltE1, resulted in a reduction in molecular weight of *all* visible products (data not shown), indicating that the vast majority contain this exon. The uncut product was ligated into vector pCR4, and 13 clones were sequenced; these represented five distinct but related sequences, each containing the 3′-most 44–61 bp of

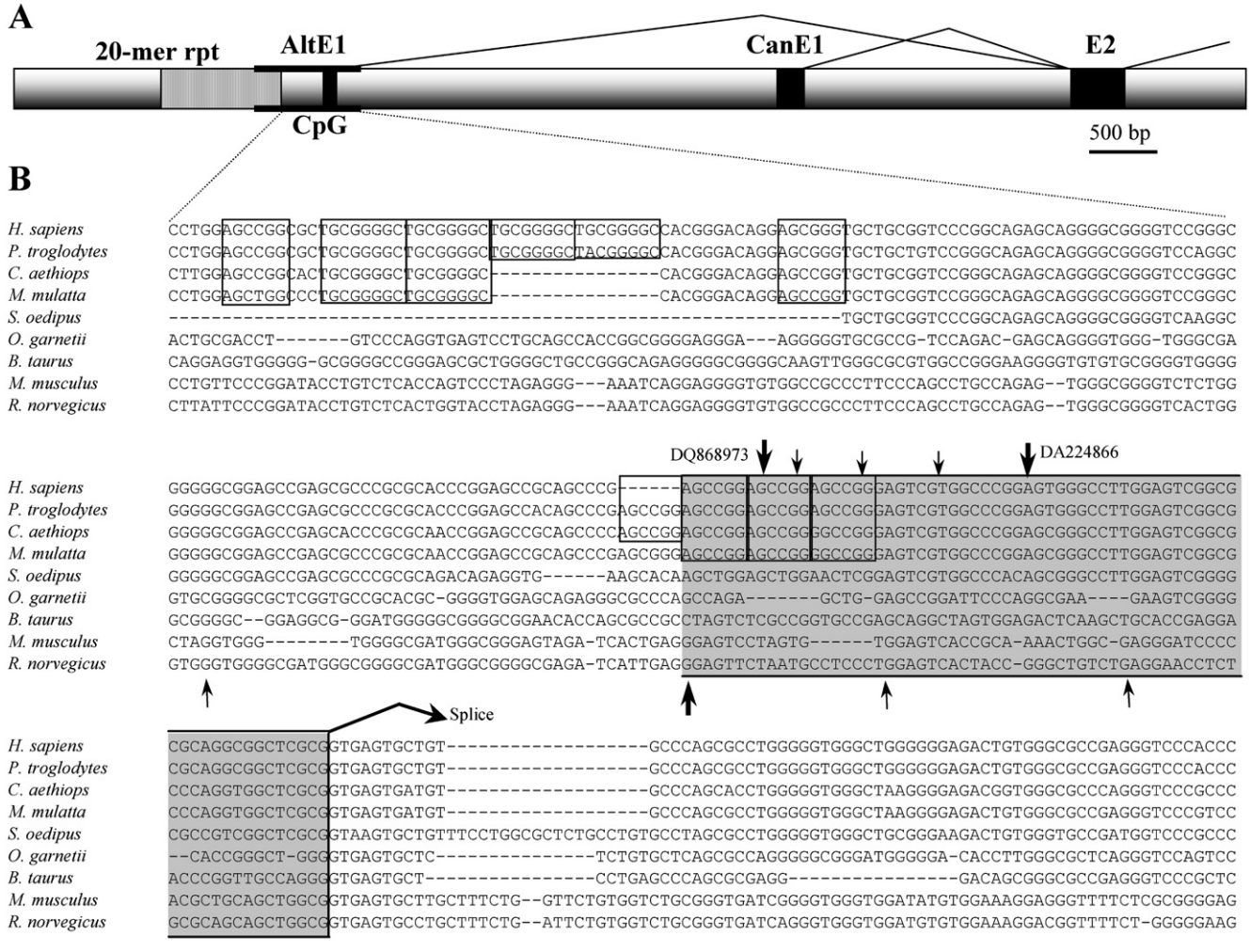

Fig. 2. Anatomy of *NOD2* alternative promoter and exon 1 (AltE1). (A) Scheme, drawn to scale, of genomic region surrounding AltE1, CanE1, and exon 2 of the human *NOD2* gene. Black boxes—exons. Vertical shading—20-mer repeat array. Bold horizontal lines—CpG island. (B) Alignment of ~300 bp surrounding AltE1 from various mammalian species. Vertical arrows above sequence show 5′ ends of 5′cRACE products obtained in this study (large arrows indicate three longest clones, DQ868973, and oligo-cap clone DA224866), while vertical arrows below sequence show 5′ ends of mouse ESTs (large arrow indicates most prevalent 5′ end). Gray box—AltE1. Small boxes—hexamer and octamer motifs described in text. Sequences acquired from Ensembl genomic sequences except for *Ce. aethiops* and *S. oedipus* from this study (Accession Nos. DQ868974andDQ868975).

AltE1 cleanly spliced onto exon 2. No clones contained sequence corresponding to CanE1. The broad distribution of 5′ ends (arrows in Fig. 2) may represent a genuinely broad spectrum of transcriptional start sites, as has been described for CpG-rich promoters [28], or may reflect stochastic stalling of reverse transcriptase while copying this highly GC-rich exon. The sequence corresponding to the longest clone is presented in Accession No. DQ868973. During the preparation of this article we noted the new release of an EST sequence, DA224866, this being the only human EST representing the 5′ end of the *NOD2* transcript; this also uses AltE1. As it was generated as part of a study of oligo-cap cDNA libraries [29] and starts close to our 5′cRACE ends, this confirms that we are likely to be identifying true transcriptional start sites rather than reverse transcriptase stalling events.

The homogeneous results of this 5′cRACE experiment suggest that AltE1 is the main constitutive first exon of *NOD2* in white blood cells. An alternative explanation is that some property of CanE1 strongly biases against its amplification in 5′ cRACE. However, while CanE1 (~170 bp) is longer than AltE1 (~60 bp), the latter is much more GC-rich (80% compared with 58%), a notorious challenge for both reverse transcriptase and *Taq* DNA polymerase. On balance, we believe that the preponderance of AltE1 over CanE1 (13 clones versus 0) probably reflects the template population of mRNAs.

We then used semiquantitative RT-PCR to compare the relative usage of the canonical and alternative *NOD2* promoters in a range of human tissues relevant to CD pathogenesis. For this, we used forward primers roughly equal distances from the 3′ ends of the two first exons, together with identical reverse primers in exon 4, giving similarly sized products from the two transcripts (745 bp for that arising from the AltE1 promoter and 757 bp for that arising from the CanE1 promoter). All primers were designed to work equally well on human and New World monkey sequences. As controls we set up RT-PCRs capable of amplifying all full-length *NOD2* transcripts: one from exon 2 to exon 4 (using the same exon 4 reverse primers as above) and one from exon 4 to exon 12. We controlled for RNA loading by amplifying transcripts from the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*). RNA samples chosen were human white blood cells, breast (the source of the original canonical *NOD2* cDNA), colon, duodenum, SW480 (a cell line derived from a human colon adenocarcinoma [30]), and white blood cells from the common marmoset, *Ca. jacchus*. The RT-PCR results show that the *NOD2* gene (as detected by reactions spanning exons 2–4 and 4–12) is transcribed at higher levels in human and marmoset white bloods cells and in the SW480 cell line and at lower levels in human breast, colon, and duodenum (see Fig. 3). The promoter usage was striking; RT-PCR products generated from transcripts originating from the alternative promoter (AltE1) mirrored the generic *NOD2* transcripts almost exactly, while those originating from the canonical promoter (CanE1) were barely amplified in SW480 and not at all in the primary tissue samples (even breast; see Fig. 3).

We take the combined results of 5′cRACE and semiquantitative RT-PCR to confirm that the alternative promoter is the overwhelmingly major source of *NOD2* transcripts in the pathogenetically relevant human tissues and that use of the canonical promoter is negligible. This tallies with the situation in mouse, in which the orthologous AltE1-like exon is used in all 35 pertinent ESTs.

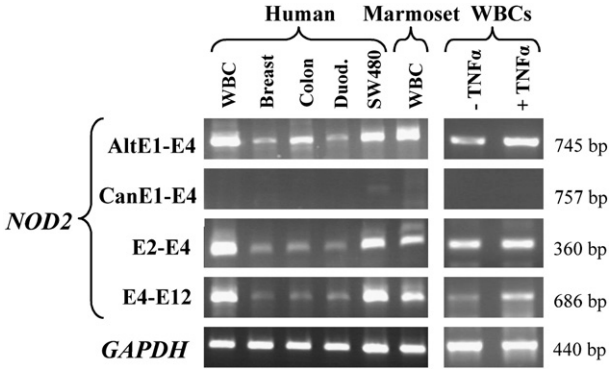

Fig. 3. *NOD2* alternative promoter is the major promoter used in clinically relevant tissues. Results of semiquantitative RT-PCR using primers in indicated exons of the *NOD2* gene (see left) are shown. Left-hand side shows a range of human samples, as labeled, together with white blood cells (WBCs) from *Ca. jacchus*. Right-hand side shows human WBCs cultured with and without TNF-α. Sizes of products are given at right.

As two previous studies [26,31] have shown substantial up-regulation of CanE1 promoter-mediated transcription following stimulation with TNF-α, we cultured human white blood cells with and without TNF-α, extracted the RNA, and performed semiquantitative RT-PCR as described above (Fig. 3, right). This resulted in no appreciable change in expression from either promoter.

The genomic context of AltE1

AltE1 lies ~3.5 kb upstream of CanE1 and ~6 kb upstream of exon 2 of the *NOD2* gene (see Fig. 2A). Like the mouse first exon, it lies in the middle of a classic CpG island (%GC=68.8%, obs/exp CpG=0.727, length 693 bp), some of which is repetitive and therefore potentially variable (4 tandem copies of the octamer TGCGGGGC in human and chimpanzee and 2 in monkeys, 3 tandem copies of the hexamer AGCCGG in human and 4 in chimpanzee and African green monkey—the latter repeats lie within the exon itself). Like many CpG islands, the region is rich in Sp1 binding sites. Unusually, the sequence ~300–1000 bp upstream of the human transcriptional start site is a highly repetitive region (see Fig. 2A and Supplementary Fig. 2), comprising approximately 32 copies of an ~20-bp GT-rich motif (GTGTGTGTGTTGGGAAGCG or similar). This sequence is present in chimpanzee, in which the array comprises only about 22 copies, and in the rhesus monkey *Ma. mulatta* (about 20 copies). The existence in GenBank of three independent genomic sequences of the human alternative *NOD2* promoter region (Accession Nos.: long allele AJ303140, short allele AC007728 and AC007608), between which this repeat array differs in length, suggested to us that there is a substantial likelihood of length polymorphism.

Table 1
Association test of *NOD2* promoter repeat region haplotypes with CD

	Haplotype					Total
	9–3–4	10–3–4	8–5–5	9–5–5	10–5–5	
Controls	283	7	1	732	11	1034
Control frequency (%)	27.4	0.68	NA	70.8	1.06	
Crohn's disease (CD)	303	7	0	738	18	1066
CD frequency (%)	28.4	0.66	NA	69.2	1.69	
p value[a]	0.62	NA	NA	0.46	0.3	

[a] *p* value for difference in haplotype frequency between cases and controls.

Association studies of variants adjacent to AltE1

AltE1 is a novel exon of the *NOD2* gene, which had not to our knowledge been screened for variants or tested for association with CD, a disease that is associated with variants in other parts of this gene. We therefore set out to assess the degree of sequence and length variation in the environs of the true human promoter and first exon. The promoter repeat array was PCR-amplified from the genomic DNA of 44 unrelated individuals (CD and ulcerative colitis patients) and sequenced; this revealed a G → A single base substitution (rs11645448 in dbSNP) and three biallelic dinucleotide repeat length polymorphisms (see Supplementary Fig. 2). No variants were seen in the exon itself, and none are listed in dbSNP. The first, second, and third copies of the 32 ~20-bp repeats were found to have alleles comprising 9 or 10, 3 or 5, and 4 or 5 GT dinucleotides, respectively. The observed haplotypes comprised alleles 9–3–4 ($n=13$), 9–5–5 ($n=74$), and 10–5–5 ($n=1$); the sequences of haplotypes 9–3–4 and 10–5–5 are shown in Supplementary Fig. 2. The A allele of SNP rs11645448 was in complete linkage disequilibrium with the 9–3–4 haplotype in our sample of 88 chromosomes.

A case–control study of association with CD was performed by using a length-based assay to genotype the polymorphic repeat region in 533 CD patients and 517 controls. Two additional aberrantly sized products were occasionally observed during this study, and sequencing revealed novel rare haplotypes with GT repeat lengths of 10–3–4 and 8–5–5. None of the above three common haplotypes showed an association with CD (χ^2 test: $p=0.58$) or linkage disequilibrium with any of the known *NOD2* mutations (see Table 1).

Discussion

The principal finding described in this paper is the identification of a novel promoter and first exon (AltE1) for an important human disease gene, *NOD2/CARD15*. We make a case that this is in fact the constitutive source of *NOD2* transcripts in blood leukocytes and colon, as opposed to the previously described canonical promoter and first exon (CanE1) [12], as follows:

(a) The alleged primate-specific ORF in CanE1 has been disrupted at least twice in primate evolution by frameshifting mutations. Analysis of those ORFs that are intact shows no evidence of the purifying selection that might be expected of a functional peptide-encoding region. The CanE1 ORF therefore seems not to be a significant biological entity.
(b) In our experiments, disruption of this ORF results in an inability to translate from the second methionine codon (Met28). This suggests that the presence of CanE1 inhibits translational initiation at Met28, perhaps because of the presence of an additional conserved upstream out-of-frame methionine codon, a factor previously shown to be inhibitory to translational initiation [32].
(c) The human CanE1 has no clear orthologue in nonprimate mammals. The mouse exon 1, which is present in all 35 relevant ESTs in GenBank, has orthologues in all mammalian genomes examined. We call the human orthologue of this sequence "alternative exon 1."
(d) AltE1 lies in a strong CpG island and its donor splice site is conserved in all mammalian sequences examined. It has no methionine codons.
(e) AltE1 is the only sequence observed upstream of *NOD2* exon 2 when 5′RACE is performed on human white blood cell RNA. The 5′ end of the 5′RACE products lies in a position similar to that of the 5′ ends of mouse *Nod2* ESTs and the sole human oligo-cap EST.
(f) RT-PCR performed using primers spanning *NOD2* AltE1 and exon 4 yields a similar expression profile across pathologically relevant tissues (white blood cells, colon) compared to that obtained from pan-*NOD2* reactions spanning exon 2 and exon 4 or exon 4 and exon 12. Primers spanning CanE1 and exon 4 yielded no product from these tissues.

We believe that together these findings make a strong case for AltE1 being the constitutive first exon of the human *NOD2* gene and, by extension, suggest that AltE1's orthologues will be the constitutive first exons of their respective genes. We found *NOD2* transcripts bearing CanE1 extremely hard to detect; indeed they were seen only in a tumor cell line and even then at levels at least an order of magnitude lower than AltE1. This suggests that CanE1 is rarely used in normal human tissues, at least in those considered important in the pathogenesis of CD (white blood cells, colon). Of course it remains possible that while AltE1 is the main promoter in normal tissues, CanE1 may become activated under certain conditions (as has been demonstrated [26,31]), such as cytokine stimulation, resulting (in humans) in an isoform 27 amino acids larger than the normal isoform.

This last point is relevant to two studies that focused on the regulation of the canonical promoter [26,31]. A previous study by some of us showed that CanE1-specific *NOD2* transcripts were up-regulated up to 20-fold by TNF-α and/or IFN-γ treatment [26], which may then allow them to approach the abundance of the AltE1-specific transcripts. The parallel study by Gutierrez et al. also showed 70-fold up-regulation by TNF-α, though in this case *all NOD2* transcripts were assayed (by using primers in exons 2 and 4) [31]. Both groups also showed that the CanE1 promoter was responsive to these agents in a luciferase-based reporter assay and that this responsiveness was dependent

on a specific NF-κB site. Clearly both of these studies were conducted before our discovery that AltE1 is the main constitutive first exon in unstimulated tissues; our preliminary findings reported here fail to detect changes in the relative usage of AltE1 and CanE1 following stimulation with TNF-α, and further work needs to be done to reconcile these apparent discrepancies.

The identification of mammalian first exons is notoriously difficult; they are often noncoding or contain little conserved coding potential, they are poorly conserved between species (with some spectacular exceptions), poorly represented in cDNA libraries, and directly obtainable only by 5′RACE, an arguably challenging technique when applied to low-abundance transcripts. As there exists a finite, albeit low, number of transcripts bearing CanE1, it is perhaps unsurprising that the ability of various workers to amplify such transcripts by RT-PCR has evidently masked the identity of the constitutive first exon of the human *NOD2* gene.

One of the immediate implications for the NOD2 field is the simple fact that the constitutive human NOD2 protein has 27 fewer amino acids than previously thought. This presents nomenclatural problems for the many known coding variants, all of which have been named, as is conventional, according to the number of the codon in the reference sequence that they affect. The reference coding sequence starts at Met1 of a cDNA sequence bearing CanE1. Our discovery means that this numbering system is formally incorrect (the old Met28 is the first codon of the principal *NOD2* reading frame) and that there is no conventional way of indicating variants in the true first exon, AltE1. Given the large amount of data concerning *NOD2/CARD15* variants that have previously been published, however, we do not believe that renumbering of the human *NOD2* reference sequence would be helpful. A further implication of our work is that primate and nonprimate NOD2 proteins are more similar than was previously thought, which should make extrapolation of findings in model organisms simpler.

Perhaps the most pressing implication of the discovery of AltE1 was that the promoter and one of the exons of this important disease gene had never to our knowledge been tested for CD-predisposing variants. We screened for variants in this region and tested a range of length-variant haplotypes for association with CD in a case–control study. No association was observed, and we conclude that common variation in the promoter and first exon region of the *NOD2* gene does not predispose to (or protect against) CD. Given the types of *NOD2* variants that *are* known to be associated with CD, this is perhaps unsurprising, and it might be expected that variants that qualitatively or quantitatively alter the transcription of the *NOD2* gene, rather than the structure of its protein product, would have distinct phenotypic consequences.

The well-known high incidence of colitis and bowel cancer in captive *S. oedipus* appears to be unlikely to be causally related to the initially intriguing frameshifting mutations in *NOD2* CanE1. However, our perusal of the literature emphasizes that while the high prevalence of bowel cancer may be restricted to *S. oedipus,* a high incidence of chronic colitis, described as "Crohn's disease-like" [24], is widespread in captive individuals of most callitrichid species. Clearly a *NOD2* mutation in the last common ancestor of callitrichids remains a candidate contributor to this colitis. Our previous study [6] highlights 57 amino acid differences between the *S. oedipus* NOD2 protein sequence and the human one; 23 of these substitutions are not present in any other known vertebrate NOD2 sequence. In addition, we have found that *S. oedipus, S. labiatus,* and *Ca. jacchus* (and presumably, therefore, other callitrichids) have a duplication of exons 5 and 6—the relationship of these copies to the NOD2 gene is unknown, though they do not appear to be incorporated into the NOD2 transcript (data not shown). All of these callitrichid-specific variants are candidate factors in the etiology of captive callitrichid colitis.

In summary, we believe that we have compiled a persuasive case that the constitutive promoter of the human *NOD2/CARD15* gene in pathologically relevant tissues has previously escaped identification. This brings the human gene more in line with other mammalian *NOD2* genes and will enable studies of the transcriptional regulation of *NOD2* to proceed on a surer basis. In addition, we show that there is no evidence for the involvement of promoter variants in the pathogenesis of CD.

Materials and methods

DNA and RNA preparation

Genomic DNA from callitrichids was previously extracted from blood or tissue using Qiagen kits (see Mundy and Kelly [33] for details of sample provenance). Fresh whole human or *Ca. jacchus* blood was diluted in 4 volumes of red blood cell lysis buffer (0.1 mM EDTA, 155 mM NH_4Cl, 10 mM $KHCO_3$) and incubated on ice for 15 min, followed by centrifugation at 400 *g* for 10 min. The resulting mononuclear cell pellet was resuspended in 2 volumes of red blood cell lysis buffer and further centrifuged at 400 *g* for 10 min at 4°C. For TNF-α stimulation, the mononuclear cells were resuspended in RPMI medium (Sigma, Poole, UK) supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 100 U/ml penicillin, and 0.1 mg/ml streptomycin and were cultured overnight at 37°C in 5% CO_2. Aliquots of 1×10^6 cells (viability >95%) were transferred to a final volume of 3 ml RPMI medium, with or without 10 μg/ml TNF-α (R&D Systems, Abingdon, UK) and cultured for a further 24 h. Mononuclear cells were harvested by centrifugation at 400 *g* for 10 min. RNA was isolated from the primary and cultured mononuclear cell pellets using RNeasy Spin columns (Qiagen, Crawley, UK) according to the manufacturer's protocol. Additionally, total RNA from ascending colon, duodenum, and adult breast tissue was purchased from Stratagene (La Jolla, CA, USA); these tissues are from three distinct single donors.

Genomic PCR, cloning, and sequence analysis

To amplify the CanE1 region from primate genomic DNA, the following primers were used: CanE1F, 5′-TGGAAGGCTGGTTGGCCAACTCTG-3′, and CanE1R, 5′-ATGTCGCGGCCAAGGATGAAAGAA-3′ (product size 314 bp). For amplification of AltE1 genomic region from *H. sapiens, Ce. aethiops,* and *S. oedipus,* a range of primers based on the human sequence was used, different pairs working on each species (primer sequences available on request; product sizes ranging up to 497 bp). Amplification reactions were performed in 10-μl reaction volumes with 20 ng DNA, 0.5 U *Taq* DNA polymerase (Promega Corp., Southampton, UK), 1 μl 10× reaction buffer, 2.5 mM $MgCl_2$, 0.2 mM dNTP, 2 μM each oligonucleotide primer. Reactions were heat-denatured for 2 min at 90°C followed by 30 cycles of 90°C for 30 s, 60°C for 30 s, 72°C for 1 min. PCR products were sequenced directly as follows: 4 μl of PCR product was incubated with 1 μl ExoSAP-IT (USB, Staufen, Germany) at 37°C for 15 min, followed by 80°C for 15 min, and sequenced

using the BigDye v3.1 dye terminator kit (Applied Biosystems, Warrington, UK) according to the manufacturer's instructions. Products were analyzed using an ABI 3730XL DNA analyzer.

Heterologous expression and Western blots

Expression constructs with differing first exon sequences were generated by synthesizing long (109-nt) primers comprising ~80 nt of human, primate, or mutant CanE1 or AltE1 followed by 23 nt of exon 2 (see Supplementary Fig. 1A). These were used for genomic PCR with a common reverse primer, which incorporated sequence encoding an HA tag. Genomic DNAs from *H. sapiens, S. oedipus,* and *A. trivirgatus* were used as templates. PCR products were cloned into mammalian expression vector pGW1-CMV, verified by sequence analysis, and transfected into cultured COS-7 cells using Polyfect (Qiagen) reagent according to manufacturer's recommendations. Cell lysates were subjected to discontinuous tricine–PAGE using 10–20% gradient gels (Bio-Rad), electroblotted onto Hybond-N nitrocellulose membrane (Amersham), and probed with primary antibodies to the HA tag (Santa Cruz Biotech) and to α-tubulin (Santa Cruz Biotech). Secondary antibody conjugated to HRP was detected using an ECL detection kit (Amersham) according to the manufacturer's recommendations.

Semiquantitative RT-PCR

One microgram of RNA and 0.5 μg random hexanucleotide primers were mixed and heated at 70°C for 5 min. Reverse transcription was then performed in a 25-μl volume at 37°C for 1 h using 1 U RNase inhibitor (Promega), 200 U MMLV reverse transcriptase (Promega), 0.5 mM dNTP, and Promega RT reaction buffer. cDNA amplification was performed on 2.5 μl of the resulting cDNA products, using 0.2 U *Taq* polymerase (Promega), Promega PCR buffer, 2 mM $MgCl_2$, 0.2 mM dNTP, and 2 μM each primer in a final volume of 25 μl. Reactions were denatured at 95°C for 2 min and then subjected to a variable number of cycles of 92°C for 30 s, 65°C for 30 s, and 72°C for 1 min, followed by 72°C for 5 min. Products were analyzed at various cycle numbers to ensure that the reactions were still in a logarithmic phase; those shown in Fig. 3 had been subjected to 40 cycles. The following primers were used to amplify parts of the *NOD2* transcript: Ex4F, 5′-CTTGGTGTCTGCAAGGCTCT-3′, and Ex12R, 5′-CAAAGCAAGAGTCTGGTGTCC-3′ (product size 686 bp); CanEx1F, 5′-CTCTCMYGATGAGGAGGAAAG-3′, and Ex4R, 5′-CCAGACCTCCAGGACATTCT-3′ (757 bp); Ex2F, 5′-ATTGTCAGGAGGCTCCACAG-3′, and Ex4R (360 bp); AltEx1F, 5′-CGGAGYGGGCCTTGGAGTCG-3′, and Ex4R (745 bp). Ambiguous nucleotides, given in IUPAC nomenclature, indicate degeneracies incorporated to allow amplification of both human and marmoset sequences. The control reaction used primers GAPDH_F, 5′-TCATCTCTGCCCCCTCTGCT-3′, and GAPDH_R, 5′-CGACGCCTGCTTCACCACCT-3′ (440 bp).

5′cRACE

5′cRACE was performed on human white blood cell RNA according to Maruyama et al. [27], using Ex2RT (5′-phospho-TCCCATGCCAGGTCCAGCATG-3′) to prime the reverse transcription, Ex2OutF (5′-CCAGCCTCTCTCCCACTTGGC-3′) and Ex2OutR (5′-CCAGGACAGCAGCCAGTCCAG-3′) to perform the first round PCR, and Ex2InF (5′-CAGAAGCTCATCGCGGCTGCC-3′) and Ex2InR (5′-CCTGAGACCAGCAGCTCGACC-3′) to perform the second round PCR. Products larger than 150 bp were gel-purified using QiaQuick columns (Qiagen) and TA-cloned into pCR4-TOPO vector (Invitrogen), after which plasmid DNA was prepared using a Qiaprep plasmid miniprep kit (Qiagen).

Genotyping of polymorphic microsatellites

PCR amplification of the *NOD2* promoter microsatellite repeat region upstream of AltE1 was performed on 10 ng of genomic DNA using 0.3 U Super Hot *Taq* DNA polymerase (Web Scientific, Crewe, UK), 0.5 μl 10×reaction buffer, 1 mM $MgCl_2$, 30 ng each oligonucleotide primer (CARDrepF, 5′-TTTTCCTAGTTTTGGCTGTGTG-3′, and CARDrepR, 5′ACACAAACATGCTTCCCAATATA-3′), 0.1 mM each dNTP, and 160 nM fluorescently labeled dUTP (Applied Biosystems) in a final volume of 5 μl. Reactions were heat-denatured for 6 min at 95°C followed by 30 cycles of 95°C for 30 s, 65°C for 30 s, and 72°C for 1 min.

For genotyping of the microsatellite repeats, 1 μl of PCR product was resuspended in 9 μl HIDI formamide (Applied Biosystems) containing the Genescan 500[ROX] size marker (Applied Biosystems) and resolved using the ABI 3730XL DNA analyzer. Genotypes were automatically called using the GeneMapper software and validated by comparison with a subset of alleles in which the microsatellite repeat length had been previously determined by DNA sequencing.

Clinical samples and statistical analysis

The ascertainment, diagnosis, and collection of samples from Crohn's disease patients have been previously described [34]. Population controls were obtained from the 1958 British Birth Cohort DNA collection (http://www.cls.ioe.ac.uk). The distribution of the upstream length variant haplotypes was compared between cases and controls using a χ^2 test. Observed data for rare haplotypes 10–3–4 and 8–5–5 were pooled to avoid violation of the assumptions of the χ^2 distribution. The frequencies of common haplotypes were also compared using binomial tests for difference in proportions.

Bioinformatics

Public domain genomic sequences were obtained from the Ensembl Genome Server (http://www.ensembl.org/index.html) and NCBI (http://www.ncbi.nlm.nih.gov), curated in VectorNTI (InforMax, Inc.), and manipulated in BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Novel sequence data were analyzed using ContigExpress (InforMax, Inc.). Alignments were done using ClustalW and most phylogenetic analysis was performed using programs from the Phylip package [35], all within the BioEdit platform. CpG island analysis was done by CpG Island Detector at http://cpgislands.usc.edu/ [36]. K_a/K_s ratios were calculated using the FUGE bioinformatics platform (http://www.bioinfo.no/tools/kaks) [37].

Acknowledgments

Marmoset blood was a kind gift from Michael Jackson at King's College London. We are indebted to the following for funding the project: K.C. Wong Scholarship (S.T.), Leverhulme Trust and BBSRC (N.I.M.), the Wellcome Trust (S.A.F., A.C.), and the Guy's and St Thomas's Charity (K.K.). We acknowledge the use of samples from the 1958 British Birth Cohort DNA Collection, funded by the Medical Research Council (Grant G0000934) and the Wellcome Trust (Grant 068545/Z/02).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.07.009.

References

[1] D.K. Podolsky, Inflammatory bowel disease, N. Engl. J. Med. 347 (2002) 417–429.
[2] J.P. Hugot, et al., Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease, Nature 411 (2001) 599–603.
[3] J. Hampe, et al., Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations, Lancet 357 (2001) 1925–1928.

[4] Y. Ogura, et al., A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease, Nature 411 (2001) 603–606.
[5] S. Lesage, et al., CARD15/NOD2 mutational analysis and genotype–phenotype correlation in 612 patients with inflammatory bowel disease, Am. J. Hum. Genet. 70 (2002) 845–857.
[6] K. King, et al., Mutation, selection, and evolution of the Crohn disease susceptibility gene CARD15, Hum. Mutat. 27 (2006) 44–54.
[7] M. Chamaillard, et al., Gene–environment interaction modulated by allelic heterogeneity in inflammatory diseases, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 3455–3460.
[8] M. Economou, T.A. Trikalinos, K.T. Loizou, E.V. Tsianos, J.P. Ioannidis, Differential effects of NOD2 variants on Crohn''s disease risk and phenotype in diverse populations: a metaanalysis, Am. J. Gastroenterol. 99 (2004) 2393–2404.
[9] S.E. Girardin, et al., Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection, J. Biol. Chem. 278 (2003) 8869–8872.
[10] N. Inohara, et al., Host recognition of bacterial muramyl dipeptide mediated through NOD2: implications for Crohn's disease, J. Biol. Chem. 278 (2003) 5509–5512.
[11] C. Miceli-Richard, et al., CARD15 mutations in Blau syndrome, Nat. Genet. 29 (2001) 19–20.
[12] Y. Ogura, et al., Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-kappaB, J. Biol. Chem. 276 (2001) 4812–4818.
[13] P. Rosenstiel, et al., A short isoform of NOD2/CARD15, NOD2-S, is an endogenous inhibitor of NOD2/receptor-interacting protein kinase 2-induced signaling pathways, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 3280–3285.
[14] E. Leung, J. Hong, A. Fraser, G.W. Krissansen, Splicing of NOD2 (CARD15) RNA transcripts, Mol. Immunol. 44 (2007) 284–294.
[15] N.K. Clapp, et al., The marmoset as a model of ulcerative colitis and colon cancer, Prog. Clin. Biol. Res. 186 (1985) 247–261.
[16] J.D. Wood, et al., Colitis and colon cancer in cotton-top tamarins (Saguinus oedipus oedipus) living wild in their natural habitat, Dig. Dis. Sci. 43 (1998) 1443–1453.
[17] C. Lushbaugh, G. Humason, N. Clapp, Histology of colitis: Saguinus oedipus oedipus and other marmosets, Dig. Dis. Sci. 30 (1985) 45S–51S.
[18] A. Purvis, A composite estimate of primate phylogeny, Philos. Trans. R. Soc. Lond., B Biol. Sci. 348 (1995) 405–421.
[19] A.W. Sainsbury, J.K. Kirkwood, E.C. Appleby, Chronic colitis in common marmosets (Callithrix jacchus) and cotton-top tamarins (Saguinus oedipus), Vet. Rec. 121 (1987) 329–330.
[20] D.T. Chalmers, L.B. Murgatroyd, P.F. Wadsworth, A survey of the pathology of marmosets (Callithrix jacchus) derived from a marmoset breeding unit, Lab. Anim. 17 (1983) 270–279.
[21] N.K. Clapp, M.L. Henke, C.C. Lushbaugh, G.L. Humason, B.L. Gangaware, Effect of various biological factors on spontaneous marmoset and tamarin colitis: a retrospective histopathologic study, Dig. Dis. Sci. 33 (1988) 1013–1019.
[22] K.M. Das, et al., Mr 40,000 human colonic epithelial protein expression in colonic mucosa and presence of circulating anti-Mr 40,000 antibodies in cotton top tamarins with spontaneous colitis, Gut 33 (1992) 48–54.
[23] A. Gozalo, E. Montoya, Mortality causes of the moustached tamarin (Saguinus mystax) in captivity, J. Med. Primatol. 21 (1992) 35–38.
[24] A. Gozalo, G.E. Dagle, E. Montoya, R.E. Weller, Spontaneous terminal ileitis resembling Crohn disease in captive tamarins, J. Med. Primatol. 31 (2002) 142–146.
[25] R. Moore, N. King, J. Alroy, Characterization of colonic cellular glycoconjugates in colitis and cancer-prone tamarins versus colitis and cancer-resistant primates, Am. J. Pathol. 131 (1988) 477–483.
[26] P. Rosenstiel, et al., TNF-alpha and IFN-gamma regulate the expression of the NOD2 (CARD15) gene in human intestinal epithelial cells, Gastroenterology 124 (2003) 1001–1009.
[27] I.N. Maruyama, T.L. Rakow, H.I. Maruyama, cRACE: a simple method for identification of the 5′ end of mRNAs, Nucleic Acids Res. 23 (1995) 3796–3797.
[28] P. Carninci, et al., Genome-wide analysis of mammalian promoter architecture and evolution, Nat. Genet. 38 (2006) 626–635.
[29] K. Kimura, et al., Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes, Genome Res. 16 (2006) 55–65.
[30] A. Leibovitz, et al., Classification of human colorectal adenocarcinoma cell lines, Cancer Res. 36 (1976) 4562–4569.
[31] O. Gutierrez, et al., Induction of Nod2 in myelomonocytic and intestinal epithelial cells via nuclear factor-kappa B activation, J. Biol. Chem. 277 (2002) 41701–41705.
[32] M. Kozak, Interpreting cDNA sequences: some insights from studies on translation, Mamm. Genome 7 (1996) 563–574.
[33] N.I. Mundy, J. Kelly, Evolution of a pigmentation gene, the melanocortin-1 receptor, in primates, Am. J. Phys. Anthropol. 121 (2003) 67–80.
[34] N.J. Prescott, et al., A nonsynonymous SNP in ATG16L1 predisposes to ileal Crohn's disease and is independent of CARD15 and IBD5, Gastroenterology 132 (2007) 1665–1671.
[35] J. Felsenstein, PHYLIP—Phylogeny Inference Package (version 3.2), Cladistics 5 (1989) 164–166.
[36] D. Takai, P.A. Jones, Comprehensive analysis of CpG islands in human chromosomes 21 and 22, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 3740–3745.
[37] D.A. Liberles, Evaluation of methods for determination of a reconstructed history of gene sequence evolution, Mol. Biol. Evol. 18 (2001) 2040–2047.